LLM-Assisted Variable Annotation using the I-ADOPT Framework

Arvin Rastegar[1], Barbara Magagna[2] and Christof Lorenz[1]

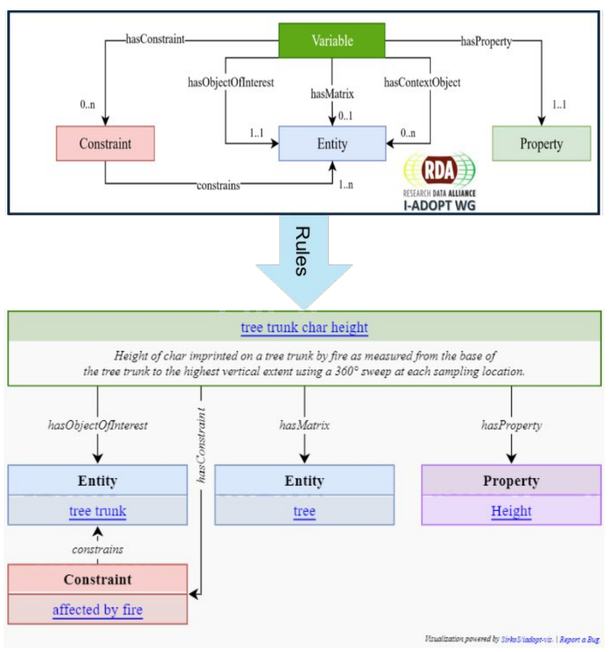
1: Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU), Garmisch-Partenkirchen, Germany
2: GO FAIR Foundation, Leiden, The Netherlands

Automating Semantic Variable Annotation for Environmental Data Interoperability

Within the Helmholtz Earth and Environment DataHub initiative, we are developing an LLM-assisted variable annotation service that leverages recent advances in Large Language Models (LLMs) to automate the semantic decomposition of variable descriptions. Our approach employs the community-driven I-ADOPT framework to break down natural-language variable definitions into essential atomic elements, ensuring naming consistency and interoperability across domains. The system incorporates retrieval-augmented generation (RAG) to access relevant literature and controlled vocabularies, enabling more precise annotations and reducing manual effort for data producers.By aligning AI-driven methods with established semantic standards, our work addresses several DSS 2025 focus areas—including Foundation Models & LLMs, AI/ML in Earth System Sciences, metadata management, and data workflows—while supporting the broader objectives of NFDI and the European Open Science Cloud. This approach enhances reproducibility, interoperability, and cross-disciplinary collaboration in day-to-day research data stewardship.

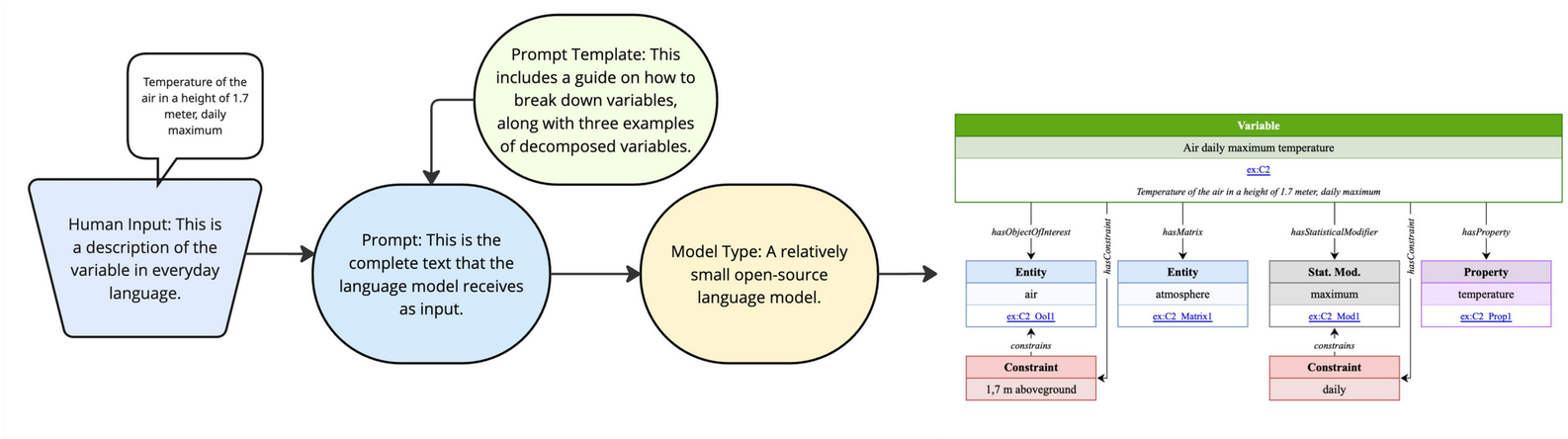
The DataHub and why consistent metadata plays a crucial role for its tools and services

Consolidation of Semantics in Metadata



The I-ADOPT framework provides the semantic foundation for mapping and consolidating controlled vocabularies by decomposing observable properties into essential atomic parts, ensuring naming consistency and interoperability.

LLM-enabled I-ADOPT: Extraction of FAIR Variable Descriptions



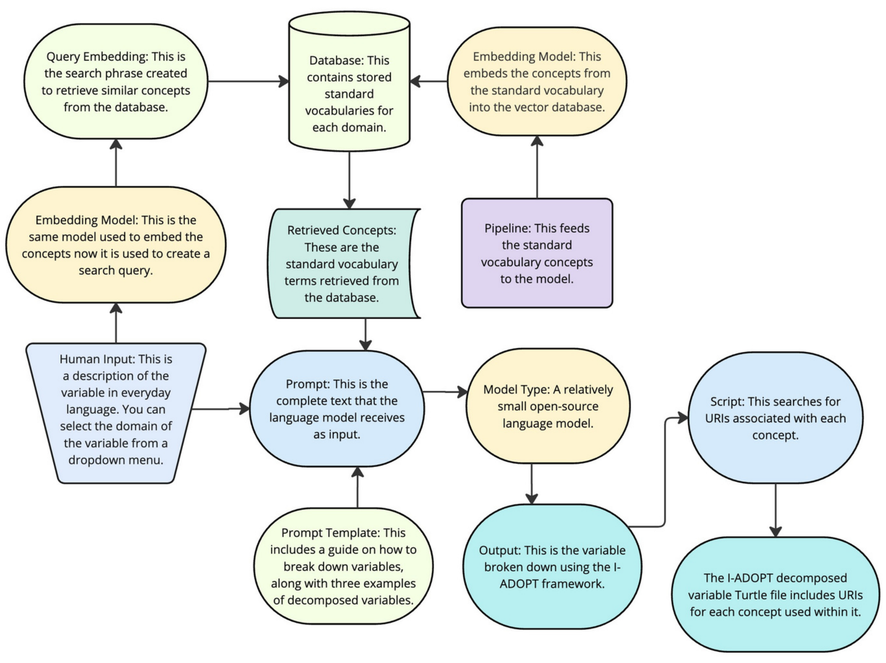
The workflow, enabled by an LLM, decomposes human-readable variable descriptions (e.g., “Temperature of the air at 1.7 m; daily maximum”) into atomic I-ADOPT elements. Guided by a prompt template with examples, the model maps each variable into structured components (entity, property, statistical modifier, constraints). This ensures machine-readable, interoperable representations across datasets, reducing manual effort, improving consistency, and supporting integration into infrastructures such as the Helmholtz DataHub, NFDI, and EOSC.

Baseline of the models

model’s name	model F1 score
qwen/qwen3-32b	0.555
meta-llama/llama-4-maverick-17b-128e-instruct	0.514
deepseek/deepseek-r1-0528-qwen3-8b	0.488
deepseek/deepseek-r1-0528	0.485

Our experiments show that simply increasing model size does not lead to major performance gains. While the larger LLaMA-4 Maverick (17B) and DeepSeek R1 (52B) models perform comparably, the more compact Qwen3-32B achieves the best F1 score (0.555). This suggests that relatively smaller open-source models can be more efficient choices for variable annotation without significant loss in accuracy.

Outlook and next steps



Using RAG to retrieve relevant literature and resources aligned with each variable as context to facilitate decomposition to I-ADOPT elements.